Characterization of X-ray data sets

Peter H. Zwart, Ralf W. Grosse-Kunsteleve & Paul D. Adams
Lawrence Berkeley National Laboratory, 1 Cyclotron Road, BLDG 64R0121, Berkeley California 94720-8118, USA – Email: PHZwart@lbl.gov; www: http://cci.lbl.gov

1. Introduction

With the emergence of structural genomics, more effort is being invested into developing methods that incorporate basic crystallographic knowledge to enhance decision making procedures (e.g. Panjikar, 2005).
A key area where some crystallographic knowledge is often vital for the smooth progress of structure solution is that of judging the quality or characteristics of an X-ray dataset. For instance, detecting the presence of anisotropic diffraction or twinning while a crystal is on the beam line, may allow the user to change the data collection strategy in order to obtain a better or a more complete data set. In post-collection analyses, the presence of (for instance) non-crystallographic translational symmetry might help the user (or program!) to solve the structure more easily.
Of course, the identification of problems is by no means a guarantee that the problems can be overcome, but knowledge of the idiosyncrasies of a given X-ray data set permits the user or software pipeline to tailor the structure solution and refinement procedures to increase the chances of success.

In this report, a number of routines are presented that assist the user in detecting specific problems or features within a given dataset. The routines are made available via the open source *CCTBX* libraries (http://cctbx.sourceforge.net) and will also be included in the next available *PHENIX* (Adams, *et al.*, 2004) release.

2. Methods

2.1. Likelihood-based scaling

Absolute scaling is performed using a maximum likelihood method as proposed by Popov & Bourenkov (2003). The X-ray amplitudes are assumed to follow a Wilson distribution, with a resolution dependent variance that takes into account the effects of geometric regularities on the average intensity (Zwart & Lamzin 2004; Morris *et al.*, 2004):

$$f(F_{obs} \mid k) = \frac{2(kF_{obs})}{\varepsilon\sigma^2(d^*)[1+\gamma(d^*)]} \exp\left[-\frac{(kF_{obs})^2}{\varepsilon\sigma^2(d^*)[1+\gamma(d^*)]} \right] \qquad \mathbf{1}$$

In the latter probability density function, $\sigma^2(\mathrm{d}^*)$ is equal to the sum of squared atomic form factors and the term $\gamma(\mathrm{d}^*)$ is a correction term accounting for resolution dependent behavior of the mean intensity due to geometric regularities. The term $\gamma(\mathrm{d}^*)$ has been obtained from 20 high quality experimental datasets in a manner similar as described by Zwart & Lamzin (2004). $\sigma^2(\mathrm{d}^*)$ is determined from the cell contents as provided by the user.

The factor ε accounts for the statistical effect of symmetry on the expected intensity (Stewart & Karle, 1976). $\mathrm{F}_{\mathrm{obs}}$ is an observed structure factor amplitude and *k* is a scale factor that brings the observation to an absolute scale with atomic displacement parameters equal to 0:

$$k = \exp[-k_s]\exp\left[-\mathbf{h}^{\mathbf{T}}\mathbf{U}^*\mathbf{h}\right] \qquad \mathbf{2}$$

The tensor $\mathbf{U}^*$ is an anisotropic atomic displacement parameter (Grosse-Kunstleve & Adams, 2002), the vector **h** is a Miller index. Note that the scalar part of the scale factor is an exponent, $\exp[-k_s]$, rather than the simple constant that is more frequently used (Giacovazzo (1992), expression 5.12). The use of an exponent has the benefit that no special precautions need to be taken during minimization procedures to ensure the positivity of *k*.

The scale factor and elements of $\mathbf{U}^*$ are determined via the minimization of the negative of a log-likelihood function:

$$\Lambda[\{F_{obs}\} \mid k_s, \mathbf{U}^*] = -\sum_{j=1}^{N_{obs}} Ln\left[f\left(F_{obs,j} \mid k(k_s,\mathbf{U}^*)\right)\right] \qquad \mathbf{3}$$

The negative log likelihood is optimized using a gradient driven L-BFGS minimizer (Liu & Nocedal, 1989). During optimization, symmetry constraints on the elements of $\mathbf{U}^*$ and its effect on the partial derivatives are taken into account (Grosse-Kunstleve *et al.*, unpublished results).

A related (and independent) implementation of the likelihood-based scaling routine is available in *PHASER* (McCoy *et al.*, 2005). An isotropic, moment based method has been implemented in ARP/wARP (Morris *et al.*, 2004).

2.2 Detection of pseudo translational symmetry

The presence of pseudo translational symmetry can often be detected by computing a native Patterson at truncated resolution. A significant off-origin peak indicates the presence of a large number of parallel inter-atomic vectors, due to translational NCS or due to an n-fold NCS axis parallel to an n-fold crystallographic axis. In order to determine whether an off-origin peak is significant, a frame of reference is needed. For this purpose, the largest off-origin peaks for roughly 500 high quality data sets from the PDB with 1 molecule in the asymmetric unit were computed and stored. In the latter calculations, only peaks further then 15 Å away from an origin peak were considered and the Patterson function was calculated using data between 10 and 5 Å resolution. The peak height of the largest peak in a Patterson map was expressed as a fraction of the height of the Patterson origin peak.

The distribution of the selected peaks heights can be described by an extreme value distribution (Weisstein, 1999). The collected set of Patterson peaks denoted by $\{Q_{\max}\}$ are limited between 0 and 1. The following standard transformation (Zwart, A.P., personal communication) scales the set of Patterson peak heights to the domain $[0,\infty)$:

$$Q_{\max}^{/} = \frac{Q_{\max}}{1 - Q_{\max}} \qquad \mathbf{4}$$

A theorem similar to the central limit theorem, suggests that the values of $Q_{\max}^{/}$ follow a Frechet distribution (Weisstein, 1999). Applying the transformation specified in equation 4 and assuming a Frechet distribution for $Q_{\max}^{/}$ results in the following cumulative distribution function of the height of the largest off-origin peak in a Patterson map:

$$F(Q_{\max}) = \exp\left[-\left(\frac{Q_{\max}}{a(1 - Q_{\max})}\right)^{-b}\right] \qquad \mathbf{5}$$

The constants a and b of this distribution function, were fitted using likelihood methods given the observed set of Patterson peak heights. The fitted constants a and b and are equal to 6.79*10^{-2} and 3.56, respectively. The observed and modeled cumulative distributions are shown in Fig 1.

The significance of an observed off-origin Patterson peak can be assessed by computing a so-called p-value: the probability that a Patterson peak of that height or larger occurs by chance. This value is equal to $1 - F(Q_{\max})$. If a threshold of 1% is chosen, all off-origin peaks with a height larger than 20% of the origin peak are considered to be significant.

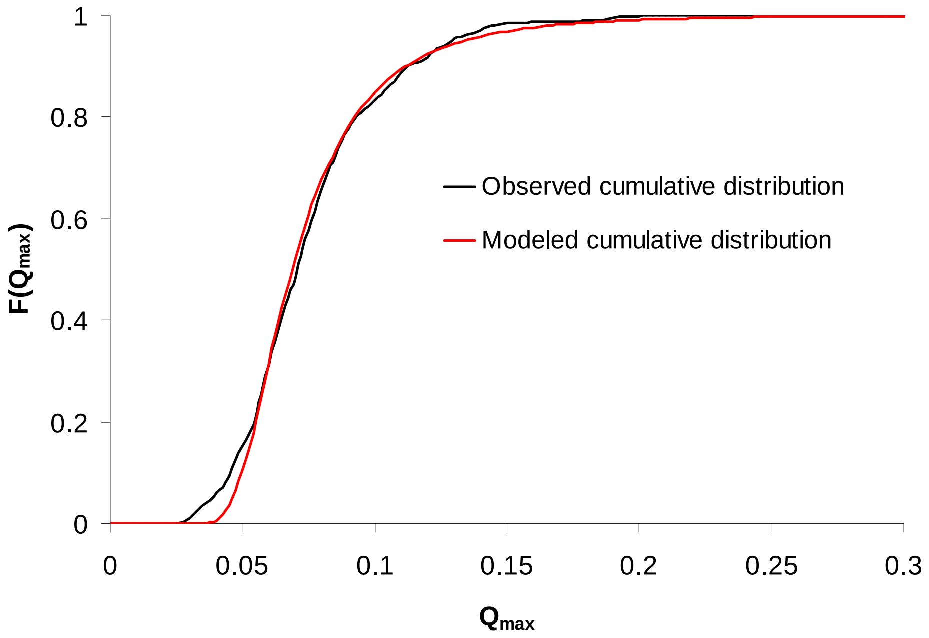

Figure 1: Observed and modeled cumulative distribution of largest off-origin Patterson peak height Q_{max}.

2.3. Twin detection

The presence of twinning can usually be identified on the basis of the Wilson or intensity ratio (e.g. Dauter, 2003). In some cases however, the presence of pseudo translational symmetry or anisotropic diffraction influences the intensity statistics in such a way that twinning cannot readily be detected, even though it is present. Therefore, the $|L|$ statistic developed by Padilla & Yeates (2003) is designed to be a more robust statistic for the detection of twinning, as it is relatively insensitive to anisotropy in the data and the presence of pseudo centering. The $|L|$ statistic is defined as follows:

$$|L| = \frac{|I_1 - I_2|}{I_1 + I_2} \qquad \mathbf{6}$$

The intensities I_1 and I_2 have associated Miller indices that are close in reciprocal space, and are not necessarily related by a twin law:

$$\mathbf{h}_1 - \mathbf{h}_2 = (d_h n_h, d_k n_k, d_l n_l) \qquad \mathbf{7}$$

d_h, d_k and d_l are random signed integers and the constant n_h,n_k,n_l are chosen on the basis of the location of significant off-origin Patterson peaks.

The first and second non-central moments of $|L|$ are equal to 1/2 and 1/3 for untwined, acentric data, respectively. If twinning is present, the moments are lowered and reach a value of 3/8 and 1/5 for perfectly twinned data. In order to detect twinning, the same data sets as used to obtain a distribution of Patterson peak heights, was used to compute <|L|> and <|L|2> values for data between 10 and 3.5 Å resolution. The resulting set (<|L|>,<|L|2>) was used in the construction of a multivariate Z-score, known as the Mahalanobis distance (Mardia, 1980). For a given observed (<|L|>,<|L|2>) pair, the Mahalanobis distance is equivalent to the distance of the given pair to the multivariate mean in units of standard deviation. Values of the Mahalanobis distance larger then 3 indicate that the (<|L|>,<|L|2>) pair is outside the range expected for experimental data sets and could thus indicate twinning.

The dependence of the Mahalanobis distance on the twin fraction is shown in Fig. 2, and indicates that X-ray data sets with a twin fraction larger then 6% have an expected Mahalanobis distance larger than 3.

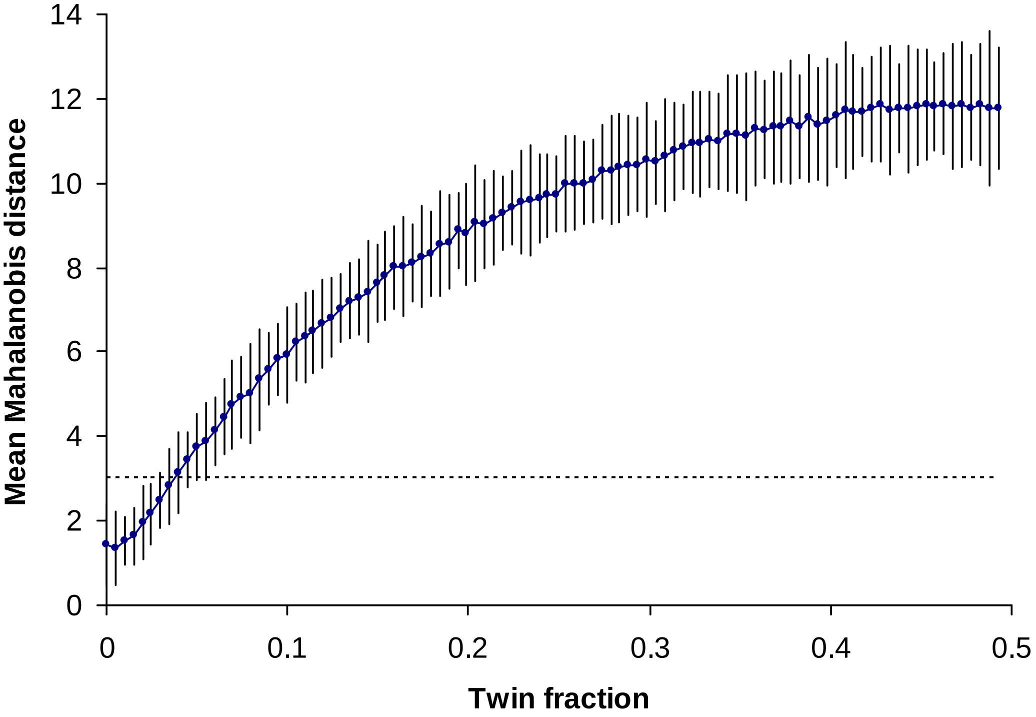

Figure 2: The expected Mahalanobis distance for the fist and second moment of |L| of an X-ray data set (blue dots). The vertical error bars span three times the estimated standard deviation of the expected Mahalanobis distance. The black dotted horizontal line is drawn for the Mahalanobis distance being equal to three. The values shown in this figure were obtained via numerical simulations.

2.4. Estimation of the twin fraction

Although twin detection and the estimation of a twin fraction are related problems, it is useful to leave these topics separated, as will become clear in section 3.3.

Estimating the twin fraction can be carried out in a number of ways. First of all the |H| test (Yeates, 1988; 1997) gives a numerically easily accessible estimate of the twin fraction. A Britton analysis (Fisher & Sweet, 1980), although less straightforward than the H-test, is another common way of estimating the twin fraction.

Another way would be to estimate the twin fraction using the |L|-statistic. As the distribution of L for a given twin fraction is known for acentric reflections, a maximum likelihood approach can be used to estimate a twin fraction. A comparison of the 3 implemented twin fraction estimation procedures is shown in Fig. 3, where the mean values of estimated twin fraction are plotted given the true twin fraction. Although results of these analyses show that the estimation of the twin fraction via the L-statistic is sub-optimal in comparison to the two other methods, especially for large twin fractions, it could be potentially be useful in cases when a two-fold non crystallographic symmetry axis is parallel to a potential twin operator. In that case, the independence between intensities required by the Britton and H-test, is violated, resulting in an overestimation of the true twin fraction. The determination of the twin fraction *via* the L-test is most likely less sensitive to these types of problems. The biggest limitation of twin fraction estimation using the L statistic is its large associated standard deviation (results not shown).

It should be noted that the distribution of the normalized intensity can also be used to estimate the twin fraction within a maximum likelihood framework (Zwart *et al.*, unpublished results). However, the drawback of this method is its extreme sensitivity to translational symmetry.

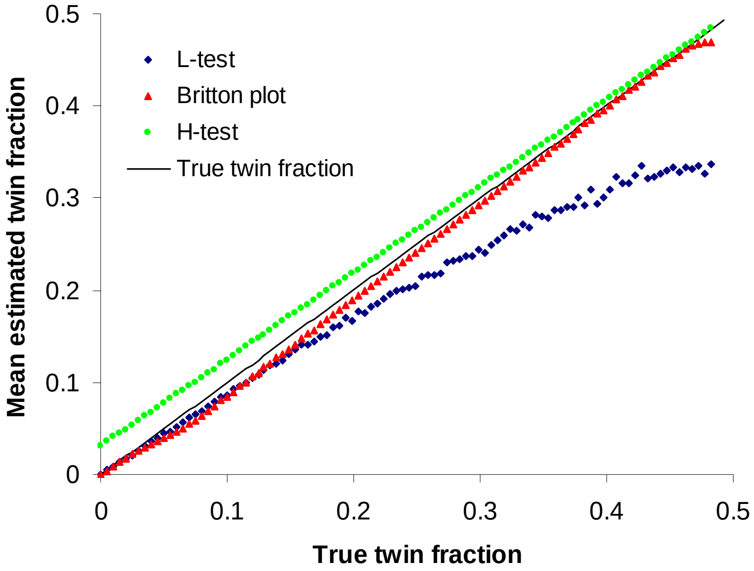

Figure 3: The estimation of the twin fraction on simulated, twinned data. Mean values over 100 trials per true twin fraction are shown. Both the H test and the Britton plot methods behave reasonably over the full range of twin fractions. The estimate of the twin fraction *via* the L statistic shows considerable bias, especially at large twin fractions.

3. Examples

3.1. The effect of anisotropy correction on the cumulative intensity distribution

The X-ray data from PDB entry 1awu is known to be anisotropic (see for instance Padilla and Yeates, Fig. 1) and the resulting cumulative normalized intensity distributions differ significantly from the theoretically expected distributions. However, as a result of the likelihood-based anisotropic scaling procedure outline above, the estimated anisotropic overall B-value can be used to correct for the observed anisotropy. The effect of the anisotropy correction on the cumulative intensity distribution is shown in Fig. 4.

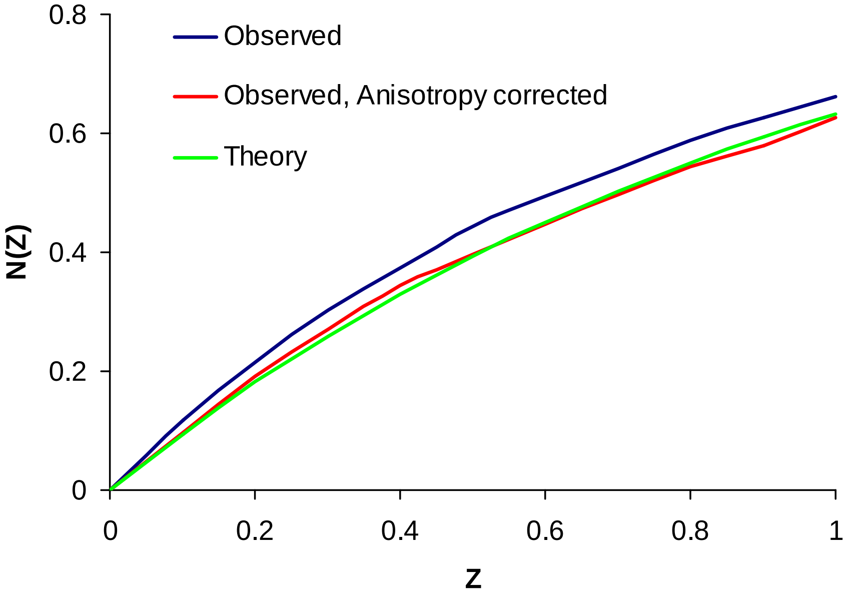

Figure 4: The effect of anisotropy correction on the cumulative intensity distribution.

3.2. Detection of non-crystallographic translational symmetry

The detection of non-crystallographic translational symmetry is illustrated using 4 example data sets obtained from the PDB. The datasets used are 1sct, 1ihr, 1c8u and 1ee2. 1sct is a classic example of pseudo centering, whereas 1ihr and 1c8u are both structures with a two-fold NCS axis (almost) parallel to a two-fold (screw) axis. 1ee2 does not possess any non-crystallographic translational symmetry.

The results for detection of translational symmetry via the presence of significant peaks in the native Patterson function are illustrated in Table 1.

Table 1: The detection of pseudo translational symmetry.

PDBID	Peak Height	p-value (%)	$<I^2>/<I>^2$	<|L|>
1sct	77%	0.0000094	2.81	0.490
1ihr	45%	0.0014	2.51	0.539
1c8u	20%	1	2.22	0.493
1ee2	10%	15	2.09	0.497

Note that the peak height (and thus the p-value) is correlated with the intensity ratio. The local intensity statistic <|L|> is however less sensitive to the presence of pseudo centering.

3.3. Detection of twinning and estimation of the twin fraction

The detection of twinning is illustrated using 5 examples obtained from the PDB. For each data set, the relevant statistics are given, as well as the reported twin fraction, if available. The twin laws for each test case were derived automatically from first principles (Flack, 1987; Grosse-Kunstleve *et al.*, 2005).

Table 2: Detection of twinning. The p-value is the p-value corresponding to the height of the largest off origin Patterson peak height. Maha(L) denotes the Mahalanobis distance of the observed (<|L|>,<|L|^2>) pair.

PDBID	Space group	Twin operator	p-value (%)	$<I^2>/<I>^2$	Maha(L)	Estimated twin fraction			Reported twin fraction*
						L-test	*Britton*	*H-test*	
1hfo	C2	h,-k,-h-l	52	2.00	0.58	0.00	0.01	0.02	None
1o0i	C2	h,-k,-h-l	28	2.09	1.08	0.00	0.44	0.46	N.A.
1hh8	C2	h,-k,-h-l	83	1.89	5.62	0.08	0.02	0.09	0
1xed	$P2_1$	h,-k,-h-l	38	1.79	6.34	0.10	0.35	0.38	0.37
1ap9	$P6_3$	h,-h-k,-l	58	1.84	7.48	0.12	0.28	0.33	None

*: None: no twinning was mentioned in the publication; N.A.: No publication available.

Although most of the test cases are easy to interpret (1hh8, 1xed and 1ap9 are all most likely twinned and 1hfo is not twinned), the X-ray data of 1o0i behaves as if it is untwinned, but intensities related by the putative twin operator are highly correlated, resulting in a estimated twin fraction of larger then 0.4. This can be rationalized by postulating that the twin operator is in fact a crystallographic symmetry element and that the reported space group is too low.

Note that if the decision about whether or not the data are twinned were made solely on the basis of the estimated twin fraction, 1o0i would be flagged as a potential perfect twin, even though the intensity statistics indicate that the structure is not twinned.

4. Conclusions

The routines presented here are aimed to provide the crystallographer with a set of statistics characterizing a given data set. The likelihood-based scaling routine provides an easy, non-graphical way of detecting anisotropy of the data by inspecting the elements of the estimated anisotropic tensor.

For the detection of pseudo translational symmetry and twinning, a similar philosophy is adopted: the summary statistics of the given data set are listed within the context of a reference set of known structures. The non-graphical nature of these analyses allows a straightforward way of incorporating general crystallographic experience into automated structure solution pipelines and allows expert and non-expert users to quickly place the results in context.

The algorithms are available as part of the open source CCTBX libraries (http://cctbx.sourceforge.net) and will also be available via future *CCP4* releases that incorporate the CCTBX.

5. Acknowledgements

We gratefully acknowledge the financial support of NIH/NIGMS through grants 5P01GM063210, 5P50GM062412 and 1R01GM071939. Our work was supported in part by the US department of Energy under Contracts No. DE-AC03-76SF00098 and DE-AC02-05CH11231.

References

Adams, P. D., Gopal, K., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy,
A. J., Moriarty, N. W., Pai, R. K., Read, R. J., Romo, T. D., Sacchettini, J. C., Sauter, N.
K., Storoni, L. C. & Terwilliger, T. C. (2004). *J. Synchrotron Rad.* **11**, 53-55.

Dauter, Z. (2003). *Acta Cryst.* **D59**, 2004-2016.

Fisher, R. G. & Sweet, R. M. (1980). *Acta Cryst.* **A36**, 755-760.

Flack, H.D. (1987). *Acta Cryst.* **A43**, 564-568.

Giacovazzo, C. (1992). *Fundementals of Crystallography*, Oxford University Press.

Grosse-Kunstleve, R.W. & Adams, P.D. (2003). *J. Appl. Cryst.* **35**, 477-480.

Grosse-Kunstleve, R.W., Afonine, P.A., Sauter, N.K. & P.D. Adams. (2005). *IUCr Computing Commission Newsletter* **5**.

Liu. D.C. & Nocedal, J. (1989). *Mathematical Programming* **45**, 503-528.

Mardia, K.V., Kent, J.T. & Bibby, J.M. (1980). Academic Press, London, UK.

McCoy, A.J., Grosse-Kunstleve, R.W., Storoni, L.C. & Read, R.J. (2005). *Acta Cryst.* **D61**, 458-464.

Morris, R.J., Blanc, E. & Bricogne, G. (2003). *Acta Cryst.* **D60**, 227-240.

Morris R.J., Zwart P.H., Cohen S., Fernandez F.J., Kakaris M., Kirillova O., Vonrhein C.,
Perrakis A. & Lamzin V.S. (2004). *J. Synchrotron Rad.* **11**, 56-59.

Padilla, J.E. & Yeates, T.O. (2003). *Acta Cryst.* **D59**, 1124-1130.

Panjikar, S., Parthasarathy, V., Lamzin, V.S., Weiss, M.S. & Tucker, P.A. (2005). *Acta Cryst.* **D61**, 449-457.

Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* **D59,** 1145-1153.

Stewart, J.W. & Karle, J. (1976). *Acta Cryst.* **A32**, 1005-1007.

Weisstein, E.W. (1999) "Extreme Value Distribution.", Mathworld: http://mathworld.wolfram.com/ExtremeValueDistribution.html

Yeates, T. O. (1988). *Acta Cryst.* **A44**, 142-144.

Yeates, T. O. (1997). *Methods Enzymol.* **276**, 344-358.

Zwart, P.H. & Lamzin, V.S. (2004). *Acta Cryst.* **D60**, 220-226.